1990

Speech intelligibility estimation via neural networks

Stephen Knight

Rochester Institute of Technology
School of Computer Science and Technology

Speech Intelligibility Estimation via Neural Networks

by
Stephen Knight

A thesis, submitted to
The Faculty of the School of Computer Science and Technology,
in partial fulfillment of the requirements for the degree of
Master of Science in Computer Science.

Approved by: _____
Professor John A. Biles

_____
Professor Dale E. Metz

_____
Professor Peter G. Anderson

Wed, Jul 18, 1990

SAMPLE statements to **authorize** or **deny** permission to reproduce an RIT thesis.

1. Title of thesis Speech Intelligibility Estimation via Neural Networks

I, Stephen Knight hereby **grant permission** to the Wallace Memorial Library of RIT to reproduce my thesis in whole or in part. Any reproduction will not be for commercial use or profit.

Date 7/30/1990

ABSTRACT

Current methods of speech intelligibility estimation rely on the subjective judgements of trained listeners. Accurate and unbiased intelligibility estimates have a number of procedural and/or methodological constraints including the necessity for large pools of listeners and a wide variety of stimulus materials. Recent research findings however, have shown a strong relationship between speech intelligibility estimates and selected acoustic speech parameters which appear to determine the intelligibility of speech. These findings suggest that such acoustic speech parameters could be used to derive computer-based speech intelligibility estimation, obviating the procedural and methodological constraints typically associated with such estimates.

The relationship between speech intelligibility estimates and acoustic speech parameters is complex and nonlinear in nature. Artificial neural networks have proven in general speech recognition that they are capable of dealing with complex and unspecified nonlinear relationships. The purpose of this study was to explore the possibility of using artificial neural networks to make speech intelligibility estimates. Sixty hearing-impaired speakers, whose measured speech intelligibility ranged from 0 to 99%, were used as subjects in this study. In addition to measuring speech intelligibility, the speech of these subjects was digitally analyzed to obtain 6 acoustic speech parameters that have been found to critically differentiate English phonemes. The subjects were divided into two sub-groups. One of the subgroups was used to train a variety of back-propagation neural networks and the other was used to test the ability of the neural networks to make accurate speaker-independent speech intelligibility estimates. The artificial neural network that seemed to be the most efficatious for making speaker-independent speech intelligibility estimates employed a bipolar squash function and scaled values of the speech parameters. Compared to listener judgements the overall accuracy of the network's speech intelligibility estimates was a respectable 83%. These findings suggest that with expanded subject populations and more acoustic speech parameters it might be possible to create a practical computer based tool capable of objectively determining speech intelligibility.

Table of Contents

Chapter 1

Introduction

While artificial neural networks have been used for many years in the field of speech recognition in an attempt to allow computers to accurately determine what a human has said, the field of speech intelligibility has been primarily using statistical analysis in an effort to develop criteria for judging how well human speech is spoken. Research into speech intelligibility has shown, however, that like speech recognition, the acoustic parameters involved in speech-intelligibilty are also associated in non-linear relationships.

This study gives an overview of some of the results of current research into speech intelligibility, the acoustic parameters that have been determined to have an effect on speech intelligibility and the results of experiments using a variety of neural networks to determine their ability to estimate speech intelligibility.

Chapter 2 provides background on speech intelligibility, the acoustic parameters involved, and the differences between the field of speech

intelligibility and speech recognition.

Chapter 3 discusses neural networks in general, the two different types of neural networks (the classification and regularity detector networks) that were used for this study, and the algorithms and features that make up a neural network's composition.

Chapter 4 elaborates regarding the implementation of the networks, that is, the acoustic variables used for input data, the composition of the training sets, and the output analysis. Modifications to the standard algorithms given in Chapter 3 are also given and the reason for these changes.

Chapters 5 and 6 provides the results, limitations, possible enhancements, and conclusions of the study. Chapter 5 presents the accuracies of the various network configurations, while Chapter 6 gives a discussion on the effects and limitations presented by the available training examples and the possible enhancements that could be done to the experiment to improve the accuracies of the neural networks used.

Chapter 2

Background

2.1 Speech and Intelligibility

For many years, speech-language pathologists have used speech intelligibility measurements as one criterion for the assessment of the severity of speech disorders, following Van Riper's long held notion that "speech is defective when it is conspicuous, unintelligible, or unpleasant." With specific regard to hearing-impaired speakers, Monsen (1981) suggested several important uses for speech intelligibility assessment, including monitoring progress in speech therapy, comparing methods of speech training, and evaluating candidates for mainstreaming. Yorkston and Beukelman (1981) recommend speech intelligibility measurement because results of intelligibility testing are easily communicated to the speaker's family and other professionals. In addition, they concluded that the strong relationship between intelligibility and information transfer suggests that speech intelligibility measures can provide a functional index of communication performance. Subtelny (1977) expressed a similar

opinion about the value of intelligibility measures for the speech of the hearing-impaired population: "For many years the difficulties and limitations in evaluating the intelligibility of deaf speech have been recognized. This fact has necessitated considerable study to establish the reliability and validity of intelligibility assessments and to define the variables influencing intelligibility".

Currently, these speech intelligibility measurements are carried out using two basic techniques. The first technique employs a rating scale to estimate intelligibility in a known context. Generally, commonly used rating scales employ an equal appearing interval scale whose scale values range from 1 to 5, 1 to 7, or 1 to 10. These numeric values are associated with antonymic definitions (ie: 1 equals "completely unintelligible", 3 equals "somewhat intelligibile", 5 equals "completely intelligible").

Rating scale procedures require that one or more trained judges listen to a speaker uttering a standard group of sentences or reading a prose passage. Then, hopefully using the same standards to judge such speech samples, the listener assigns a numeric score that corresponds to his perception of the speakers intelligibility. Rating-scale procedures have two well documented procedural constraints;

> <u>experience</u> - persons accustomed to listening to a particular type of disordered speech (like the speech of the deaf) exhibit

a 10% to 14% advantage in message comprehension over naive listeners (Monsen, 1978).

individual idiosyncrasies - the application of idiosyncratic judgements despite similar training which violates the assumptions of the rating-scale procedures.

In addition to these attendant procedural complications, Samar & Metz (1988) have recently demonstrated that at least one popular speech intelligibility rating-scale procedure exhibits gross violations of measurement prediction in the midrange of speech intelligibility (ie: rating levels 2 through 4).

The second method of estimating speech intelligibility is a verbatim transcription of a standard group of linguistically equated sentences. Each judge is instructed to write down verbatim what he hears, and the number of correctly identified words is expressed as a percentage score, which is taken as the speaker's intelligibility level. This verbatim write-down procedure limits idiosyncratic judgements from influencing the overall intelligibility score, and as such, possesses an advantage over rating-scale procedures. Additionally, the write-down procedure has good face validity and is frequently the procedure of choice in many research and clinical applications. However, write-down procedures have a major procedural constraint that requires that listeners not hear the same speech material more than once because of familiarity effects. As such,

one needs a large corpus of linguistically equated stimulus materials and/or a large pool of listeners.

Perhaps the most important limitation of both these intelligibility estimation procedures is that the resultant estimation (i.e.; numerical scaled value or percentage score) possesses no explanatory power. That is, a numerical rating and/or percentage score does not convey any information regarding the underlying nature of the intelligibility deficit. This is particularly problematic with the speech intelligibility estimation of hearing-impaired speakers. As Monsen (1978) states, "When the quality of a hearing-impaired child's speech is poorer than normal, it is typically poorer than normal in a great variety of ways. In many or most cases it is difficult for even a highly trained observer to extract the source of a speech error--that is, the real acoustic reason" . Obviously, as the level of intelligibility drops, it becomes more and more difficult for a listener to derive the actual reasons that a word, or group of words, does not sound as it "should." A possible solution to the lack of specificity and explanatory power of currently employed speech intelligibility ratings may be found by isolating the acoustic features of speech that are critical to linguistic differentiation among the speech sounds. If one could quantify systematic relationships between aberrantly produced acoustic

features and speech intelligibility, it might be possible to develop a system that formally relates acoustic aspects of speech and overall intelligibility.

2.2 Potential Phonetic Acoustic Features as Intelligibility Determiners

Monsen (1978), characterizes human speech as "a complicated, coarticulated code" where the data can "vary continuously along many different dimensions." These variations have certain features that distinguish one type of sound from another and exhibit a certain amount of consistency between one speaker and another, which implies the possibility of revealing the dimensions of speech that relate to intelligibility.

An example of a phonetic feature in the time domain which is a linguistic determiner of a certain group of phonemes is Voice Onset Time (VOT). VOT is defined as the amount of time between a stop's (such as /b/ or /p/) articulatory release, and the point at which voicing onset occurs (onset of vocal fold operation). Figure 1 shows examples of the VOT when the utterances 'pat' and 'bat' are spoken by the same person.

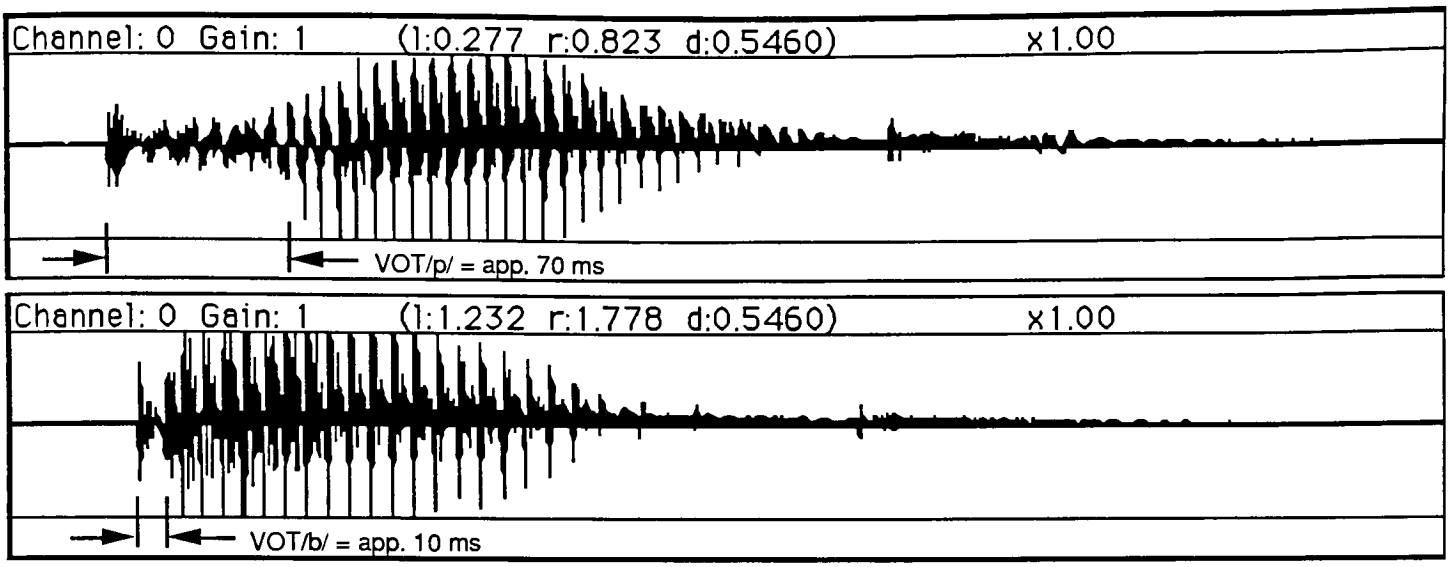


Figure 1: The above speech signals are the words 'pat' and 'bat' uttered by a human male, digitized at 10,000 Hz. The top waveform shows 'pat' and the second is 'bat'. The distance between the start of the word and where voicing occurs is the Voice-Onset-Time (VOT) and is marked for both waveforms.

Notice that the Voice Onset Time for /p/ is approximately 70 ms., whereas the VOT for /b/ is approximately 10 ms. This temporal distinction serves as the unique perceptual cue for the linguistic separation of these two cognate sounds. As this temporal distinction decreases in magnitude, the phonemic differentiation between /p/ and /b/ is lost, and there will be an associated loss of intelligibility of some unknown magnitude.

Features in the frequency domain are far more complicated to differentiate. This is primarily due to the fact that most of the time the human voice does not emit sound at a single frequency but instead occupies a spectrum from approximately 30 to 5000 Hz. Depending on the

phoneme uttered, the frequencies involved can be just at the "high" end of the spectrum (fricatives such as /s/ and /f/), more towards the "middle" (fricatives such as /th/ and /sh/), or "high" frequency "noise" coupled to the "low" frequencies associated with voicing (fricatives such as /z/ and /zh/, for example).

The critical acoustic determiners of distinct vowel (/a/, /e/, /u/, etc.) productions is the separation of concentrations of acoustic energy known as formants. The transfer function of the human vocal tract produces six formants, which span the frequency range from about 100 Hz to 5000 Hz. In between these concentrations of acoustic energy, or formants, there is relatively little acoustic information due to the attenuating characteristics of the vocal tract. The first two formants from an adult male speaker (F_1 range equals approximately 270 to 730 Hz; F_2 range equals approximately 840 to 2290 Hz) vary systematically depending on vocal tract configuration (ie: tongue, posture, lip rounding, etc.) and serve to linguistically differentiate the vowel sounds of English.

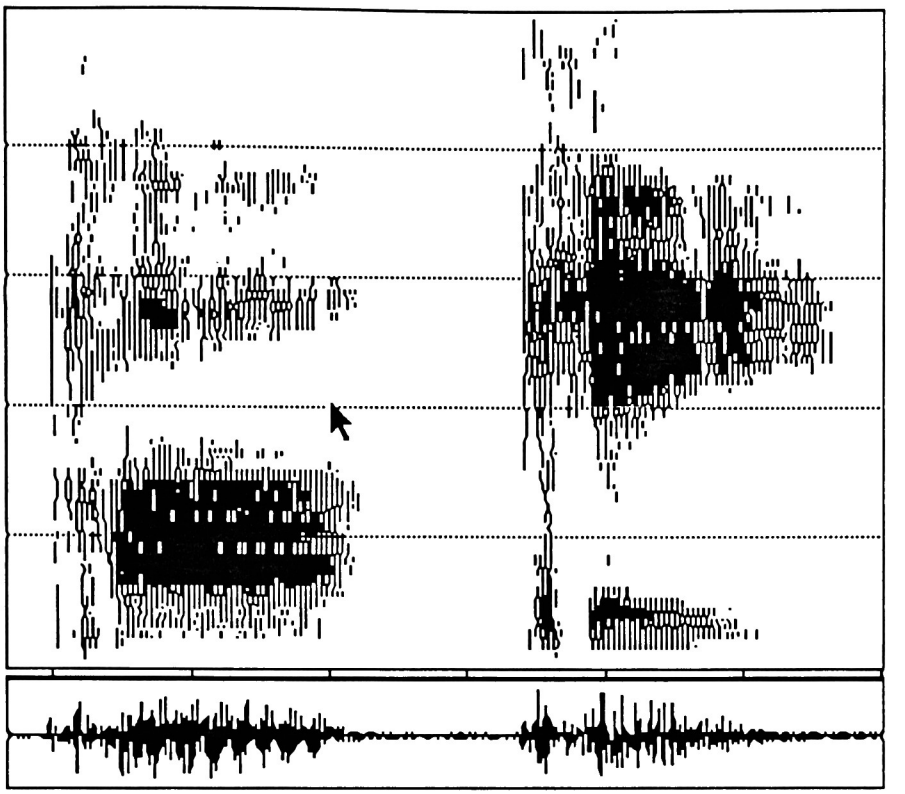


Figure 2: A spectrogram of the syllables 'pa' and 'pi' spoken by the same individual. The lower box is the time domain display of the utterances, while the upper box shows the relative magnitudes of the frequency band of 0 to 5,000Hz. The dark bands (more visible as bands in 'pi') are the formants.

Observe the separation between first and second formants in the vowels displayed in Figure 2. It is the magnitude of separation between the first and second formant that linguistically differentiates the three vowels. As this frequency separation between the two formants decreases in magnitude the phonemic differentiation between the vowels is lost and there will be an associated loss of intelligibility of some unknown magnitude.

Research into the phonetic features of speech relating to

intelligibility has yielded some acoustic features that have been shown to statistically relate to hearing-impaired speaker's intelligibility. Three speech characteristics (voice onset time difference between /t/ and /d/, the second formant difference between /i/ and /ɔ/, and a rating of the spectrographic quality of liquids and nasals) have been found to account for 73% (Monson, 1978) of the variance in speech intelligibility, as assessed by normally hearing listeners using a write-down procedure. Additionally, Weismer et al. (1988) showed that the speech intelligibility of dysarthric persons could be predicted using formant trajectories.

Using regression and principal components analysis, Metz et al. (1985) demonstrated a strong relationship between certain groupings of acoustic speech features and normally-hearing listeners' intelligibility assessments of the speech produced by 20 hearing-impaired speakers. The strong predictive relationship of these acoustic variables to speech intelligibility, and the clear convergence of the findings with previous research (Monsen, 1978; Parkhurst & Levitt, 1978), suggested that the derived factor structure reflected major underlying dimensions of speech production that have significant consequences for the intelligibility of speech produced by hearing-impaired speakers. Metz et al (1985) found

that 83% of intelligibility variance could be determined from the following weighted acoustic parameters:

1) M VOT /p/ - VOT /b/
2) M VOT /t/ - VOT /d/
3) M VOT /k/ - VOT /g/
4) M F_2 /i/ - F_2 /ɔ/
5) M F_1 /a/ - F_1 /i/
6) M F_2 change in /ai/

Parameters '1', '2', and '3', are the mean differences in VOT between the stops /p/ and /d/, /t/ and /d/, and /k/ and /g/, respectively. Parameter '4' is the mean difference between the second formants (F_2) of the the vowels /i/ and /ɔ/ (as in 'bee' and 'paw'). Parameter '5' is the mean difference between the first formants (F_1) of the vowels /a/ (as in 'father') and /i/ (as in 'bee'), while parameter '6' denotes the mean F_2 change in /ai/ (as in 'pie'). It is intuitively appealing that as the magnitudes of these six variables approaches zero, the phonemic differentiation would be lost with an attendant loss of intelligibility of some unknown magnitude.

When attempting to use the same analysis for entirely new speakers (Metz et al., 1989) the results were less encouraging. Only 74% of the variance was accounted for, and there was an associated loss of approximately 8% in efficiency when predicting intelligibility. The disparity between the results of these two studies was apparently related

to subsequently discovered non-linear relationships between the acoustic predictor variables and the speech intelligibility estimates. Such non-linearities pose major problems for factor based linear regression models. . . . Important information was lost by imposing linear model constraints on the predictive process, given that critical aspects of the acoustic signal that serve to differentiate phonemes appear to vary in a non-linear manner with speech intelligibility.

Although the non-linearities present in the features relating to speech intelligibility appear to prohibit the development of a "universal prediction formula," research into speech recognition has developed techniques that offer the possibility of speech intelligibility estimation. These techniques also offer the possibility of an appropriate explanation as to why an utterance by one speaker is more intelligible than the same utterance by another speaker.

2.3 Speech Intelligibility vs. Speech Recognition

Computer speech intelligibility estimation differs from computer speech recognition in a very fundamental way: where computer speech recognition's goal is to have an automated system listen to human speech and understand *what* a speaker said, the goal of a speech intelligibility-

rating system is to determine how *well* the speaker said it. Ideally, the speech recognition system would be able to discern the meaning of the utterances no matter what the speaker said (essentially an unlimited vocabulary). A speech intelligibility-rating system, on the other hand, should be able to provide an accurate estimate that describes the intelligibility defects in objective terms and have few procedural constraints. Ideally, the rating system also would be able to have an unlimited vocabulary, but the basic objective of the system is to determine *how well formed* a word is uttered, which means the system needs to already know what word is to be spoken. To expect the system to first decide *which* word has been spoken and then determine how *well* it was said runs up against the fact that, typically, the rating system would be used with poor speakers. As the intelligibility levels of the various speakers dropped, the recognition system's difficulty in deducing which word was spoken, thereby allowing an intelligibility rating to be given, would have a corresponding rise.

Still, there are many similarities in the information both speech intelligibility-rating and speech recognition systems need in order to perform their functions. In both cases, the starting point for solving their respective problems lies in human speech and the information contained in

it. Just as speech intelligibility phonetic features have proven to be non-linear in function, "normal" human speech also has been found to be non-linear. Research into speech recognition has developed techniques to deal with these non-linearities using various techniques such as Hidden-Markov Modeling, Boltzmann Machine Algorithms, and neural networks.

Neural networks, in various forms have been shown by Bengio et al. (1988) and Waibel et al. (1988) to offer the better possibilities for speech intelligibility estimation, due to greater accuracies.

Chapter 3

Neural Networks and Speech

3.1 Neural Networks

Two types of neural networks that have the possibility of providing some form of classification or rating of speech intelligibility are the "classification paradigm" and the "regularity detector" (Rumelhart, 1986).

In the first case, the "classification paradigm," the network is presented with a series of stimulus patterns with the appropriate category that each stimulus belongs to. The training algorithm's goal is such that when completed, the network will (hopefully) be able to correctly classify not only one of the training stimuli but also a slightly distorted version.

The second form of network appropriate for the problem of determining speech intelligibility is the "regularity detector." In this case, the network is presented with a group of stimulus patterns and the associated probability. The training is supposed to allow the network to discover the crucial features of the training population. If the training

population is appropriate, and the network topology correct, then the network should be able to produce the correct probability when relevant stimuli are presented.

In both forms, the network is a back-propagation network, consisting of an input layer (which receives the stimuli), one or more hidden layers, and an output layer. The "classification" network will have an output layer consisting of one node for each of the possible categories, while the "detector" network will have a single output node.

3.2 Networks used in Speech Recognition

While neural networks have not been used in determining speech intelligibility per se, they have been used, with varying degrees of success, in the the area of speech recognition in general. On a speaker-dependent basis, neural networks are capable of 95% accuracy in identifying the speech syllables [ba], [bi], [bu], [da], [di], [du], [ga], [gi], [gu] (Elman et al., 1988). Using a three layer, back-propagation network, Elman demonstrated a network capable of identifying whole syllables with an average of 16% errors. When the network was trained to label vowels, the error was approximately 1.5%, and training for consonants left misidentification at 7.9%.

In the above case, the network was working with unaltered input

consisting of the results of twenty 64-point FFT computations with the results of each FFT compressed down to 16 points representing the spectral magnitudes evenly spaced across the frequency range. When the input was altered with random distortion the error rates decreased to 10% for syllables, 0.3% for vowels, and 5.0% for consonants. Using this technique, a recognizer based on vowels and consonants would have a accuracy of 95%.

Training for the network consisted of random presentations of the training set (a single speaker uttering /ba/, /bi/, /bu/, /da/, /di/, /du/, /ga/, /gi/, and /gu/ fifty-six times for each syllable. A member of the training set was presented to the network, which then propagated through the layers. The error for each output unit then was backpropagated through the network so that the weights for each node could be adjusted accordingly.

At no time was the network informed as to which features were important to identifying the speech. However, later analysis of the network activity found that different hidden nodes were active only under certain conditions. For example, a hidden unit became associated with a subset of sound types such as /a/ or /i/, while no hidden unit was found to represent the /u/ sound alone. In another experiment, one hidden unit was

found to be always and **only** on for the alveolar stops (Da, Di, Du), while another was found to be the same for velar stops.

Kohonen (1988) was also able to produce a neural network capable of 96 to 98 percent accuracy for isolated-word recognition with a 1000 word vocabulary. His input consisted of FFT vectors reformatted to give spectral magnitudes across 15 frequency ranges across a bandwidth of 0 to 5 kHz. As in Elman et al. (1988), training consisted of presenting these inputs with no specification as to how the data were to be grouped. The output consisted of information detailing which word had been identified.

Using Back-propagation and Boltzmann-Machine neural networks, Bengio et al. (1988) were able to produce accuracies between 91.7% and 95.8% in recognizing the place-of-articulation in vowels. Using 144 speech samples from vowel patterns used in the continuous speech from 28 speakers, the two networks were trained with 72 tokens and tested with the other 72 tokens. Bengio et al. (1988), like Kohonen, also presented Fast-Fourier Transformed versions of the tokens. Waibel et al. (1989) also demonstrated the abilities of neural networks in determining non-English consonant recognition with accuracy rates varying from 96.6% to 100% using a vocabulary database of 5240 common Japanese words

spoken in isolation by one male native Japanese speaker. Using modular design techniques, Waibel et al. (1989) were able to produce a recognition score of 94.7% for an all-phoneme network.

3.3 Proposed Neural Network For Speech Intelligibility Estimation

As was stated earlier, Monsen (1978) and Metz et al. (1985) have demonstrated that it is possible to extract information from the acoustic speech signal and to statistically relate aspects of those signals to overall speech intelligibility. However, the apparent non-linear relationships among acoustic predictor variables and rate of speech intelligibility poses major problems for linear statistical models.

Artificial neural networks appear to have the potential of obviating many of the procedural constraints of speech intelligibility estimation and clearly have the capacity to deal with non-linear data sets. Additionally, artificial neural networks are particularly well suited for speech intelligibility estimation research due to the reasonably accurate description of the speech intelligibility domain at the acoustic level and the independent measures of speech intelligibility. In this regard, one can train the neural network to recognize and appropriately weight the acoustic variables to conform to independent listener judgements of

speech intelligibility, rather than forcing a fit through linear statistical models.

A back-propagation neural network, similar to the networks used by Elman et al. (1988) and Bengio et al. (1988), offers the possibility of increasing the accuracy of speech intelligibility estimation using acoustic variables like those proposed in Metz et al. (1985). Multi-layer back-propagation neural networks have been demonstrated (Elman, 1988) to be able to deal with complex nonlinearities occurring in a data set when the relationships between the variables are unknown.

For a back-propagation neural network to be able to deal with these non-linearities, the network requires, at minimum, an input layer and two computational layers of nodes (Lippmann,1987). The first layer consists of a layer of nodes with the number of nodes equal to the number of input parameters. The last layer in the network consists of either the category nodes, where each node represents a particular category that the input data belongs to or has been evaluated to, or a single node, whose output represents the probability of the input pattern in relation to the training data. Any layers situated between the first and last layers, if they exist, are the hidden layers. If the hidden layers have sufficient nodes for the problem at hand, they provide the network with the ability to describe

complex non-linear regions that describe the training data in such a form that when given a previously unseen pattern, the network is able to correctly set the output nodes to relate the new pattern's relationship to the previous training patterns.

Each node can be considered a structure consisting of the following parts: a series of input weighting values, a bias value, an error value, and two series of momentum values. The node sets the output value to

$$O_{node} = \text{squash}(\text{bias}_{node} + \Sigma_{j=1 \text{ to } k} W_j I_j)$$

where O_{node} is the output of the node, the squash function is a sigmoidal function capable of translating the given value between two limits, bias_{node} is a bias value for the node determined by previous training, and $(\Sigma_{j=1 \text{ to } k} W_j I_j)$ is the sum of the products of the appropriate weighting value and its corresponding input value. The weighting, momentum, and bias values are set during the back-propagation training.

The back-propagation training algorithm consists of a series of cycles, with each tick in the cycle representing a point in time where the values of the nodes in each layer can change. Before training is initiated, weights and biases in the network are set to random values between -1.0 and 1.0. When training a network consisting of an input layer and two computational layers of nodes, I, H, and O (k, l, and m signify the number

of nodes in each respective layer).

At:

time t=0: an input pattern ($x_1 \ldots x_k$), and its corresponding output pattern ($y_1 \ldots y_l$), are randomly selected from the set of input/output patterns to be learned.

time t=1: the outputs of the input layer nodes are set equal to the input pattern; $I_1 = x_1$, $I_2 = x_2$, ..., $I_k = x_k$.

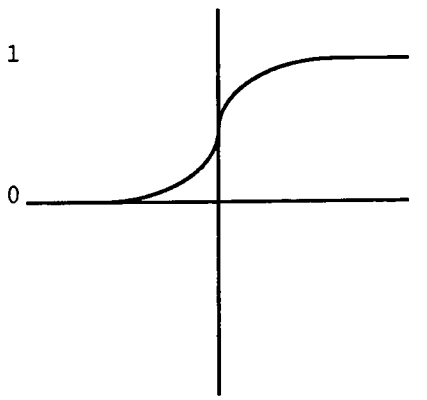
time t=2: the output activations of the hidden nodes are set to

$$H_i(t+1) = \text{squash}(\text{bias}_{Hi} + \Sigma_{j=1 \text{ to } k} W_{Hj} I_j(t))$$

where

$$\text{squash}(x) = (1 + e^{-x})^{-1}$$

for a function that will force the network to a unipolar mode, i.e.,

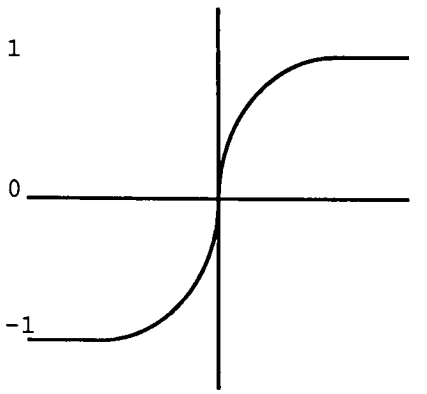


$(1 + e^{-x})^{-1}$

or, to have the network operate in a bipolar mode:

$$\text{squash}(x) = \tanh(x)$$

which results in outputs varying to x as



time t=3: the outputs of the output layer nodes are set to

$$O_i(t+1) = \text{squash}(\text{bias}_{Oi} + \Sigma_{j=1 \text{ to } k} W_{Oj} H_j(t))$$

time t=4: the resulting pattern at the output layer then becomes the network's output, and is subtracted from the target pattern $(y_1 \ldots y_l)$ to produce an error pattern. The error pattern is used to adjust the weights and biases of the output nodes using

$$\partial_{Oi} = (y_i - O_i)\, O_i\, (1 - O_i)$$

$$\Delta \text{bias}_{Oi} = \mu\, \partial_{Oi}$$

$$W_{Oij}(t+1) = W_{Oij}(t) + \mu\, \partial_{Oi}\, I_j + z(W_{Oij}(t) - W_{Oij}(t-1))$$

μ = learning rate constant

z = momentum factor

The error propagates back to the hidden layer at

time t=5:

$$\partial_{Hi} = \mu H_i (1 - H_i) \Sigma_{j=1 \text{ to } m} \partial_{Oi} W_{ij}$$

$$\Delta bias_{Hi} = \mu \partial_{Hi}$$

$$W_{Hij}(t+1) = W_{Hij}(t) + \mu \partial_{Hi} I_j + z(W_{Hij}(t)-W_{Hij}(t - 1))$$

μ = learning rate constant

z = momentum factor

The learning rate constant 'μ' determines the degree of effect errors have on the network during the training cycle. As μ increases in value, the effect the calculated error has on changing the networks bias and weight values increases. The momentum constant 'z', on the other hand, determines the effect previous weight changes will have on the current weight change. This momentum term provides a degree of smoothing that might not otherwise exist.

3.4 Network Architecture

The basic back-propagation neural network used in this thesis to deal with the non-linear complexities involved in estimating speech intelligibility is the two computational "classification" network shown in

Figures 3 and 4.

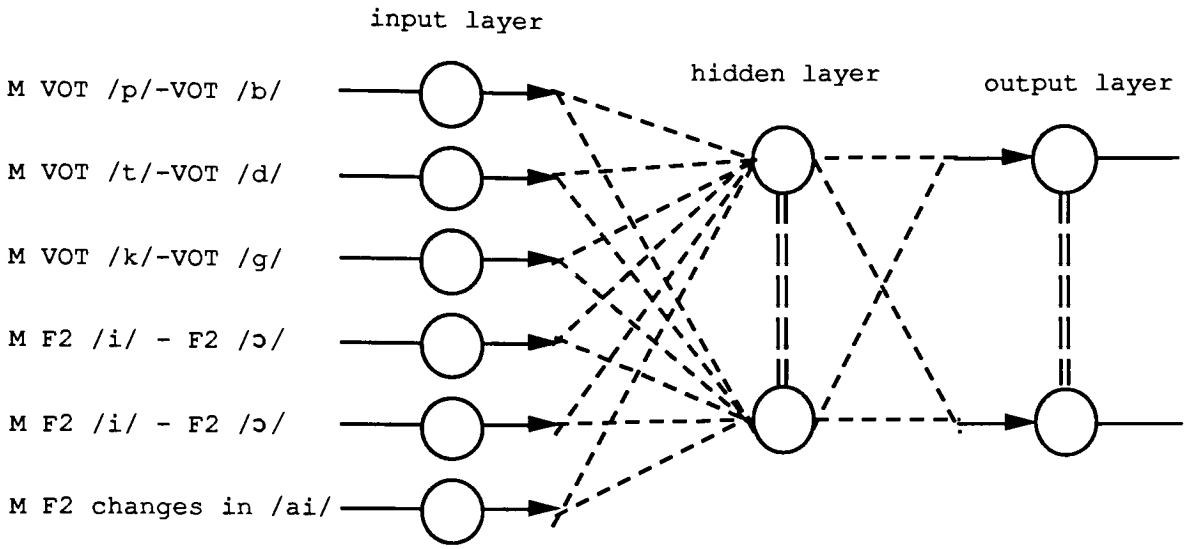


Figure 3: A two computational-layer classification network with 6 input nodes, a varying number of nodes in the hidden layer and a varying number of nodes in the output layer.

The pattern presented to the input layer consists of the values described in Metz, et. al. (1985): the average differences between VOT /p/ and VOT/b/, VOT /t/ and VOT /d/, VOT /k/ and VOT /g/, F2 /i/ and F2 /ɔ/, F2 /i/ and F2 /ɔ/, and the average F2 changes in /ai/. The output layer consists of varying numbers of nodes, with each output node representing a different "class." Which class the input pattern belongs to is derived by using the values assigned to the originating speaker to determine that speaker's intelligibility by using either a rating-scale procedure or a verbatim write-down procedure. Training the network consists of gathering a set of patterns, along with their appropriate class, and presenting the patterns in a random order. When the network has either

made the appropriate number of passes through the training set or completed a given number of "ticks," the network then is presented with a set of unfamiliar patterns for it to classify. The network's classification is shown by a single output layer node being "on" (1) with all other output-layer nodes being "off" (0).

The three-layer computational "classification" layer network shown in Figure 4 is identical to the two-layer network except that there hidden layers between the input and output layers.

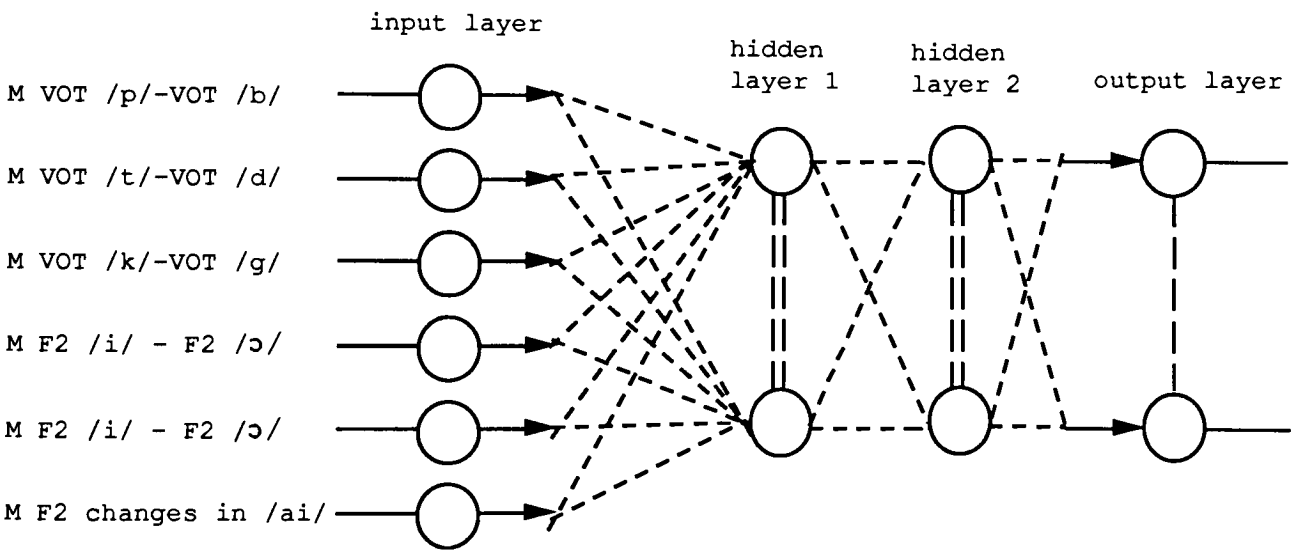


Figure 4: A two computational-layer classification network with 6 input nodes, a varying number of nodes in the hidden layer and a varying number of nodes in the output layer.

The second type of back-propagation neural network capable of estimating speech intelligibility is the "regularity detector" network (shown in Figure 5). It is different from the "classification" network in that instead of having an output node for each possible class, the

"regularity detector" has a single node for the entire output layer. When the "regularity detector" network is trained, the desired output value consists of a real number between 0 and 1 (as opposed to the "classification" network where the output nodes are either "on" (1) or "off" (0)). After the training is completed, the network should be able to respond to an unfamiliar pattern being presented to it by giving a result between 0 and 1 that represents the position that pattern would have in relationship to the patterns in the training set.

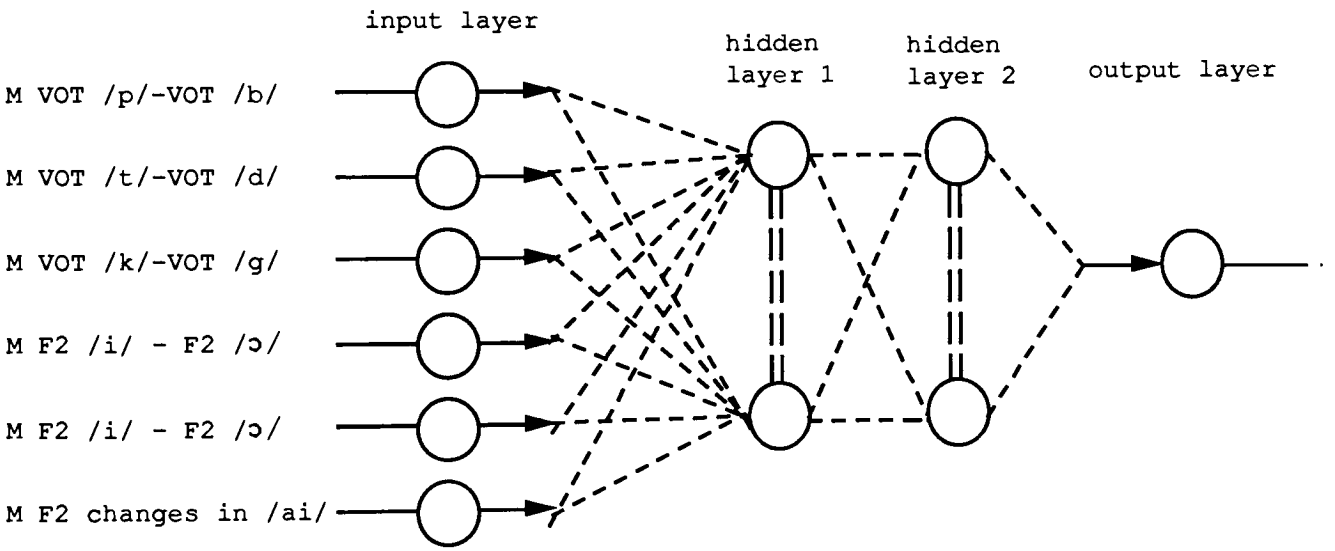


Figure 5: A three-layer "regularity detector" network with 6 input nodes, a varying number of nodes in the hidden layer and a single node in the output layer.

Chapter 4

Implementation

4.1 Acoustic Variables and Populations

In this study, the subject examples presented to the neural networks were divided into two groups; the first group being the subject data from Metz et al. (1985) and the second group being the subject data from Metz et al. (1989). When a network was trained on the first group, the second group was used as an accuracy check and vice-versa. Each individual example consisted of the six acoustic feature measurements (predictor variables) discussed earlier:

1) M VOT /p/ - VOT /b/
2) M VOT /t/ - VOT /d/
3) M VOT /k/ - VOT /g/
4) M F_2 /i/ - F_2 /ɔ/
5) M F_1 /a/ - F_1 /i/
6) M F_2 change in /ai/

and a seventh measurement, the measured intelligibility level (criteria variable). The purpose of dividing into two groups was to provide two different populations to crosscheck the result of network training. Since the purpose of this thesis was to determine if a neural network is capable

of differentiating between different levels of intelligibility on a speaker-independent basis, testing a network's accuracy by presenting the second "unknown" population was intended to give an indication of the actual degree of speaker-independence that had been achieved.

4.2 Classification Network Training

When a classification network was being used, the training and check groups were arranged in ascending order, determined by the intelligibility level. Once arranged, the groups were then divided into quintiles. Each intelligibility level then was replaced by a number 1 through 5, where 1 was 0% to 20% intelligible, 2 was 21% to 40% intelligible, etc.

When a two-class classification network was used, each class of the training population then was presented with every other class for the network to learn. Once the training had proceeded through a set number of presentation-passes or learning-ticks, the network then was tested for accuracy by presenting the "unknown" population presented to the network for its determination as to which class each of the unfamiliar individuals belonged. The size of the hidden layer(s) was varied from 4 hidden nodes to 20 hidden nodes per layer. When a five-class network was used, the entire population of each group was used as the training examples and the entire population of the "unknown" group was used for testing the

network's accuracy.

4.3 Regularity Detector Training

When a regularity detector network was being used, the same 6 acoustic variablesthat were used in the classification networks were presented . Instead of having the network's objective be to classify the example into a given group, however, the network was trained to give the actual speech intelligibility value for each example. For each training example presented to the network, the network was trained until the output node had a value within ±10% of the desired result. The training examples were presented to the network in the same groupings used in the classification networks (quintiles).

4.4 General Experimental Procedures

The general experimental procedure for evaluating the ability of neural networks to estimate speech intelligibility from a set of acoustic parameters was as follows. Two different network types were employed; classification networks and regularity detectors. Within each of these networks, a series of trials consisting of training and evaluation was conducted by varying the squash functions, the network architecture and scaling of the input data. In particular, the network architecture was modified by using two or three computational-layers with varying

numbers of hidden nodes (ie; 4 to 20 in this case). The squash functions used were the unipolar and bipolar functions discussed earlier. The input data were presented to the networks either in raw form, where there were both positive and negative variable values, in hard-limited form, where the input data was filtered to eliminate negative numbers, in log form, where the hard-limited data set was logarithmically transformed, or scaled between 0 and 0.95 relative to the maximum value of the particular input variable. All possible combinations of the above variations in network type, architecture, and data scaling were examined resulting in 816 unique networks (2 network types * 2 different hidden layers * 2 squash functions * 6 data transformations * 17 different amounts of nodes per hidden layer).

Examination of preliminary results from training several of the different networks showed a problem with the output node error calculations discussed earlier. Specifically, the equation $\partial_{Oi} = (y_i - O_i)\, O_i\, (1 - O_i)$ disrupted the training cycle by miscalculating the amount of error at a particular output node under certain conditions. If $y_i \neq O_i$ and $O_i = 1.0$, then ∂_{Oi} for that node would be set to 0.0, resulting in the network entering an endless loop in the training cycle. To avert this problem, the

error-calculating equation for the output nodes was altered to $\partial_{Oi} = (y_i - O_i)$, which allowed the training algorithm to adjust the networks according to their training results and also avoid the possible errors.

When examining the results from a 2-class classification network, accuracies of 50% were discarded. This is due to the fact that examination of the output values from these networks showed that a 50% accuracy could be obtained by the network by giving only one output value for all examples, and, therefore, in no way could the network be said to have "understood" the problem.

Examination of results from the regularity detector trials also showed occasions where the output value had been locked to a single value. These results from the regularity detector networks were also discarded if the output values were found to have been locked. If the network had more than one output value over the course of the accuracy presentations, the accuracy result for that particular check was retained.

Chapter 5

Results

5.1 Regularity Detector Results

Table 1 shows that average accuracy of the best results from the regularity detector networks ranged from 1.25% to 17.50% overall accuracy. The best accuracies from the training by quintiles ranged from 0.00% accuracy to 50% accuracy, which was the result when the network was trained with the second (2) and fifth (5) quintiles. The best accuracy of the regularity detector network when trained on the entire range of intelligibility levels was 17.50%, for a three layer network.

5.2 Classification-network Results

Overall accuracies of the various manipulations ranged from 25.00% (table 2), using hard-limited unscaled data with a unipolar squash function on two layer networks, to 83.75% (table 9), using scaled values with a bipolar squash function on a series of three layer networks. These results were obtained by averaging the accuracies of the best networks for each of the manipulations. The unscaled two layer networks ranged in

accuracy from 25.00% to 80.0% (tables 2 & 3), with the bipolar squash function networks being more accurate than the unipolar squash function networks, while the three layer networks using unscaled data ranged in accuracy from 38.75% to 82.5% (tables 4 & 5) with the bipolar squash functions again being more accurate than the networks using the unipolar squash function. The networks using the scaled data had overall accuracies ranging from 33.75% (table 6), using two layers, a unipolar squash function and the logarithm of the scaled data, to an overall accuracy of 83.75% (table 7), using three layers, a bipolar squash function and the unmanipulated scaled values. The two layer networks overall accuracies ranged in value from 33.75% to 78.12% (tables 6 & 7) and the three layer networks ranged in accuracy from 48% to 83.75% (tables 8 & 9).

The most accurate group of networks were the networks using unmanipulated scaled values, the bipolar squash function and three layers. The best accuracies obtained across each combination of classifications ranged from 56.25% to 100% (table 9) accuracy. In 80% of the combinations (8 out of 10), the speaker-independent accuracy improved when the networks were trained using the larger population group as the training examples. The accuracy of predicting the correct classification

improved as the separation between classes increased. For example, the networks were better at predicting the classification when choosing between 1 and 4 or 1 and 5 than when choosing between 1 and 2 or 2 and 3. The networks were 100% accurate when predicting class 1 and 4 speakers, 1 and 5 speakers, 2 and 4 speakers and 2 and 5 speakers.

The networks trained for all five classifications had accuracies ranging from 10% to 50.0% (table 10). The lowest accuracy of 10% occurred with scaled or unscaled, hard-limited data using three layer unipolar network and with the logarithms of scaled data, using a two layer network with a unipolar squash function. Unlike the two classification networks used earlier, the scaled, bipolar, three layer networks did not display superior accuracy to the other networks when dealing with all five classifications.

Overall, the three computational layer networks using a bipolar squash function were the most accurate regardless of whether the data was scaled or not. The networks using the scaled data demonstrated the greatest levels of accuracy and also matched or increased their accuracies when presented with the larger training group.

Chapter 6

Discussion and Conclusions

6.1 Discussion

The results given above show that artificial neural networks are capable of estimating speech intelligibility. Using the same data as Metz et al. (1985) the neural networks were almost as accurate in their ability to predict speech intelligibility (82% in this study as opposed to 83% in Metz et al. (1985)) and using the larger population data from Metz et al. (1989) the artificial neural networks were definitely superior, to the statistical methods used in that study (83% vs. 74%).

The inability of the regularity detector neural networks to accurately predict speech intelligibility (the best accuracy was 17.5%) and the inability of the classification networks to maintain as high a level of accuracy when trained to classify examples that were close in intelligibility (for example, 1 & 2) can primarily be attributed to the dearth of training examples. Other studies, such as Bengio et al. (1988) and Elman et al. (1988), had far larger training sets (144 and 505 tokens,

respectively), while this study was limited to a maximum training population of 40 tokens. Sorting the training population by the intelligibility level and dividing into quintiles produced divisions such that an intelligibility level of 0.19 could be in classification 1 while an intelligibility level of 0.21 could be in classification 2. Such a division coupled with an inadequate supply of examples could explain the classification-network's inability to accurately predict speech intelligibility, when confronted with such an environment.

A secondary problem also could have been the degree of randomization possible in presenting training examples to the network. Since neural networks are known to be altered internally when presented with a set of training examples in a different order than from that used in a previous training session (Elman et al. (1988)), it is possible that the relative lack of randomization, inherent in the present shortage of examples, could adversely effect the networks ability to learn the proper criteria for judging the training set, resulting in a larger degree of error than if a larger training set had been available.

6.2 Conclusion

While this study has demonstrated the ability of artificial neural networks to predict speech intelligibility with an equal or greater

precision than statistical analysis, there are several refinements that offer the possibility of increased accuracy.

Expanding the size of the training groups could offer an immediate improvement in results, due to the greater number of examples in each range. The problems discussed earlier, such as lack of differentiation between classes and lack of random presentation order, would most likely disappear, given a complete enough set of training examples.

Providing more acoustic information in the form of more input variables and changing the nature of the current input variables also offers the possibility of increasing the neural network's accuracy. The use of average differences between acoustic variables in this study, instead of the actual differences themselves, could have deprived the neural networks' of necessary information in each training example. Unnecessary input information possibly could be eliminated from later training by examining the internal weights and biases of the individual nodes in the computational layers. If an input variable is given no weighting at any of the hidden nodes using it as input, then it could safely be assumed that the input variable in question is unnecessary to the training being conducted.

While the results of this study could be considered limited in the

generalities that can be derived from them, overall the results do show that artificial neural networks offer the possibility of estimating speech intelligibility with a high degree of accuracy.

References

Bengio Y., De Mori R. (1988). Use Of Neural Networks For The Recognition Of Place Of Articulation.

Elman J. L., Zipser D. (1988). Learning The Hidden Structure Of Speech. *Journal Of The Acoustical Society Of America,* Vol. 83, No. 4, 1615-1626, April 1988.

Kohonen, T. (1988). The "Neural" Phonetic Typewriter. *Computer,* 11-22, March 1988

Lippmann R. P. (1987). An Introduction to Computing With Neural Nets. *IEEE ASSP Magazine,* 4-21, April 1987.

Metz, D. E., Samar V. J., Schiavetti N., Sitler R. W., Whitehead R. L. (1985). Acoustic Dimensions Of Hearing-Impaired Speakers' Intelligibility. *Journal Of Speech And Hearing Research,* Volume 28, 345-355, September 1985.

Metz D. E., Schiavetti N., Samar V. J., Sitler R. W. (1989). Acoustic Dimensions Of Hearing-Impaired Speakers' Intelligibility: Toward Automated Speech Intelligibility Assessments. *Journal Of Speech And Hearing Research,* in publication.

Monsen, R. B. (1978). Toward measuring how well hearing-impaired children speak. *Journal of Speech and Hearing Research*, 21, 197-219.

Samar V. J., Metz D. E. (1988). Criterion Validity of Speech Intelligibility Rating-Scale Procedures for the Hearing-Impaired Population. *Journal Of Speech And Hearing Research,* Volume 31, 307-316, September 1988.

Subtelny J. (1977). Assessment of speech with implications for training. In F. Bess (Ed.), *Childhood deafness* (pp. 183-194). New York: Grune & Stratton.

Parkhurst B. G., Levitt M. (1978). The Effect Of Selected Prosodic Errors On The Intelligibility of Deaf Speech. *The Journal of Communication Disorders,*Vol. 11, 249-256.

Rumelhart D. E., Hinton G. E., Williams R. J. (1986). *Parallel Distributed Procesing: Explorations in the Microstructure of Cognition. Vol. 1: Foundations.* MIT Press, 160-161.

Waibel A., Hanazawa T., Hinton G., Shikano K., Lang K. (1988). Phoneme Recognition: Neural Networks vs. Hidden Markov Models. *1988 International Conference on Acousitcs, Speech and Signal Processing,* Volume 1, 107-10.

Waibel A., Sawai H., Shikano K. (1989). Consonant Recognition by Modular Construction of Large Phonemic Time-Delay Neural Networks. *1989 International Conference on Acousitcs, Speech and Signal Processing,* Volume 1, 112-15.

Watson C. S., Reed D. J., Kewley-Port D., Maki D. (1989). The Indiana Speech Training Aid (ISTRA) I: Comparisons Between Human And Computer-Based Evaluation Of Speech Quality. *Journal Of Speech And Hearing Research,* Volume 32, 245-251, June 1989.

Weismer G., Kent R. D., Hodge M., Martin R. (1988). The Acoustic Signature For Intelligibility Test Words. *Journal Of The Acoustical Society Of America,* Volume 84, No. 4, 1281-1291, October 1988.

Yorkson K. M., Beukelman D. R. (1981). Communication efficiency of dysarthric speakers as measured by sentence intelligibility and speaking rate. *Journal of Speech and Hearing Disorders,* 46, 296-301.

	2 layers		3 layers	
training group	1	2	1	2
comparison group	2	1	2	1
1 & 2	0.1250	0.1250	0.0000	0.0000
1 & 3	0.0625	0.1250	0.0000	0.0000
1 & 4	0.0000	0.0000	0.0000	0.1250
1 & 5	0.3750	0.0000	0.0000	0.0000
2 & 3	0.1250	0.1250	0.1250	0.0000
2 & 4	0.0000	0.0000	0.0000	0.0000
2 & 5	0.3125	0.5000	0.0000	0.0000
3 & 4	0.0000	0.3750	0.0000	0.0000
3 & 5	0.0000	0.0000	0.0000	0.0000
4 & 5	0.4375	0.5000	0.0000	0.0000
average=	14.38%	17.50%	1.25%	1.25%
entire range	17.50%	0.00%	0.00%	0.00%

Table 1: Best accuracies of regularity detector networks using
a) scaled, unmanipulated input value
b) bipolar squash function
c) training for real intelligibility level.

	logarithmic		hard-limited		unmodified	
training group	1	2	1	2	1	2
comparison group	2	1	2	1	2	1
1 & 2	0.4375	0	0.5625	0	0.5625	0.625
1 & 3	0	0	0	0	0.4375	0.625
1 & 4	0.8125	1	0.625	0	0.5625	0.375
1 & 5	1	1	0	0	0.6875	0
2 & 3	0	0.625	0.5625	0	0.5625	0.75
2 & 4	0.9375	0.875	0	0.875	0.4375	0.875
2 & 5	0.9375	1	0	0	0	0
3 & 4	0.75	0.75	0.4375	0.375	0.4375	0.75
3 & 5	0.75	0	0	0.625	0.5625	0.875
4 & 5	0.4375	0	0.4375	0.625	0.4375	0.375
average=	60.63%	52.50%	26.25%	25.00%	46.88%	52.50%

Table 2: The best accuracies using
a) 2 computational layer networks
b) unipolar squash function
c) unscaled data, in logarithmic, hard-limited, and unmodified values

	logarithmic		hard-limited		unmodified	
training group	1	2	1	2	1	2
comparison group	2	1	2	1	2	1
1 & 2	0.625	0.75	0.625	0.75	0.625	0.625
1 & 3	0.6875	0.625	0.5625	0.625	0.625	0.625
1 & 4	0.875	1	0.625	0.625	0.625	0.875
1 & 5	0.9375	0.75	0.6875	0.875	0.75	0.875
2 & 3	0.75	0.625	0.625	0.625	0.5625	0.75
2 & 4	0.9375	0.875	0.4375	0.875	0.625	0.75
2 & 5	0.9375	1	0.75	0.875	0.75	0.875
3 & 4	0.75	0.75	0.4375	0.375	0	0.75
3 & 5	0.8125	1	0.5625	0.625	0.625	0.625
4 & 5	0.5625	0.625	0.5625	0.625	0.5625	0.375
average=	78.75%	80.00%	58.75%	68.75%	57.50%	71.25%

Table 3: The best accuracies using
a) 2 computational layer networks
b) bipolar squash function
c) unscaled data, in logarithmic, hard-limited, and unmodified values

	logarithmic		hard-limited		unmodified	
training group	1	2	1	2	1	2
comparison group	2	1	2	1	2	1
1 & 2	0.4375	0.25	0.625	0	0.5625	0.625
1 & 3	0.6875	0	0.815	0	0.4375	0.625
1 & 4	0.8125	0.875	0.9375	0.875	0.625	0
1 & 5	1	0.75	1	0.75	0.0526	0.75
2 & 3	0	0.4375	0.75	0.75	0.75	0.5625
2 & 4	0.9375	0.875	1	0.875	0.4375	0.375
2 & 5	0.9375	1	0.9375	0.875	0	0
3 & 4	0.75	0.625	0.75	0.125	0.4375	0.125
3 & 5	0.75	1	0.8125	0.75	0.375	0.4375
4 & 5	0.5625	0	0.4375	0.625	0.5625	0.375
average=	68.75%	58.13%	80.65%	56.25%	42.40%	38.75%

Table 4: The best accuracies using
a) 3 computational layer networks
b) unipolar squash function
c) unscaled data, in logarithmic, hard-limited, and unmodified values

	logarithmic		hard-limited		unmodified	
training group	1	2	1	2	1	2
comparison group	2	1	2	1	2	1
1 & 2	0.625	0.625	0.625	0.625	0.625	0.625
1 & 3	0.75	0.625	0.6875	0.375	0.625	0.625
1 & 4	1	0.875	0.625	0.625	0.625	0.625
1 & 5	0.9375	1	0.75	0.75	0.6875	0.875
2 & 3	0.625	0.6875	0.5625	0.625	0.5625	0.5625
2 & 4	0.9375	0.875	0.75	0.575	0.6875	0.75
2 & 5	0.9375	1	0.75	0.75	0.75	0.75
3 & 4	0.75	1	0.4375	0.4375	0	0.375
3 & 5	1	0.875	0.625	0.75	0.75	0.5625
4 & 5	0.6875	0.625	0.5625	0.375	0.5625	0.625
average=	82.50%	81.88%	63.75%	58.88%	58.75%	63.75%

Table 5: The best accuracies using
a) 3 computational layer networks
b) bipolar squash function
c) unscaled data, in logarithmic, hard-limited, and unmodified values

	logarithmic		hard-limited		unmodified	
training group	1	2	1	2	1	2
comparison group	2	1	2	1	2	1
1 & 2	0.5625	0.625	0.5625	0.375	0.5625	0.625
1 & 3	0.4375	0.25	0.75	0.625	0.8125	0.625
1 & 4	0.4375	0	0.9375	0.75	0.9375	0.875
1 & 5	0.5625	0.375	1	0.75	1	0.875
2 & 3	0.4375	0	0.625	0.75	0.625	0.75
2 & 4	0.625	0.625	1	0.875	1	0.875
2 & 5	0.4375	0.625	0.9375	0.875	1	0.875
3 & 4	0.375	0.125	0.5625	0.125	0.5625	0.125
3 & 5	0.5625	0.625	0.75	0	0.75	0.125
4 & 5	0.6875	0.125	0.4375	0.375	0.4375	0.375
average=	51.25%	33.75%	75.63%	55.00%	76.88%	61.25%

Table 6: The best accuracies using
a) 2 computational layer networks
b) unipolar squash function
c) scaled data, in logarithmic, hard-limited, and unmodified values

	logarithmic		hard-limited		unmodified	
training group	1	2	1	2	1	2
comparison group	2	1	2	1	2	1
1 & 2	0.625	0.625	0.5625	0.75	0.625	0.375
1 & 3	0.5625	0.625	0.6875	0.625	0.8125	0.75
1 & 4	0.75	0.625	1	0.875	0.9375	1
1 & 5	0.4375	0.75	1	0.875	0.9375	1
2 & 3	0.4375	0.75	0.5625	0.75	0.75	0.75
2 & 4	0.75	0.75	0.9375	0.875	1	0.875
2 & 5	0.6875	0.375	0.9375	0.875	0.9375	1
3 & 4	0.5625	0.625	0.75	0.75	0.75	0.75
3 & 5	0.625	0.75	0.75	1	0.75	0.875
4 & 5	0.625	0.75	0.3125	0.375	0.3125	0.375
average=	60.63%	66.25%	75.00%	77.50%	78.13%	77.50%

Table 7: The best accuracies using
a) 2 computational layer networks
b) bipolar squash function
c) scaled data, in logarithmic, hard-limited, and unmodified values

	logarithmic		hard-limited		unmodified	
training group	1	2	1	2	1	2
comparison group	2	1	2	1	2	1
1 & 2	0.5625	0.375	0.625	0	0.5625	0.25
1 & 3	0.5625	0.75	0.8125	0	0.6875	0.625
1 & 4	0.4375	0.625	0.9375	0.875	0.875	0.875
1 & 5	0.5625	0.375	1	0.75	1	1
2 & 3	0.4375	0.4375	0.75	0.75	0.5625	0.625
2 & 4	0.375	0	1	0.875	1	1
2 & 5	0.5625	0	0.9375	0.875	0.9375	0.875
3 & 4	0.375	0.375	0.75	0.125	0.75	0.125
3 & 5	0.375	0.625	0.8125	0.75	0.75	0.75
4 & 5	0.5625	0.375	0.4375	0.625	0.4375	0.375
average=	48.13%	39.38%	80.63%	56.25%	75.63%	65.00%

Table 8: The best accuracies using
a) 3 computational layer networks
b) unipolar squash function
c) scaled data, in logarithmic, hard-limited, and unmodified values

	logarithmic		hard-limited		unmodified	
training group	1	2	1	2	1	2
comparison group	2	1	2	1	2	1
1 & 2	0.5625	0.75	0.625	0.625	0.625	0.625
1 & 3	0.5625	0.625	0.875	0.625	0.8125	0.75
1 & 4	0.625	0.375	1	1	1	1
1 & 5	0.625	0.75	0.9375	0.875	1	1
2 & 3	0.375	0.625	0.75	0.875	0.75	0.75
2 & 4	0.75	0.625	1	1	0.9375	1
2 & 5	0.625	0.625	1	1	0.9375	1
3 & 4	0.4375	0.875	0.75	0.75	0.75	0.875
3 & 5	0.75	0.75	0.75	1	0.875	0.75
4 & 5	0.625	0.75	0.4375	0.375	0.5625	0.625
average=	59.38%	67.50%	81.25%	81.25%	82.50%	83.75%

Table 9: The best accuracies using
a) 3 computational layer networks
b) bipolar squash function
c) scaled data, in logarithmic, hard-limited, and unmodified values

		logarithmic		hard-limited		unmodified	
	training group	1	2	1	2	1	2
data type	network type						
unscaled							
	unipolar 2 layer	0.275	0.300	0.250	0.200	0.275	0.500
	bipolar 2 layer	0.325	0.350	0.250	0.300	0.300	0.250
	unipolar 3 layer	0.225	0.200	0.200	0.100	0.225	0.200
	bipolar 3 layer	0.350	0.400	0.275	0.250	0.275	0.250
scaled							
	unipolar 2 layer	0.250	0.100	0.250	0.250	0.275	0.250
	bipolar 2 layer	0.225	0.250	0.350	0.450	0.375	0.300
	unipolar 3 layer	0.225	0.200	0.200	0.100	0.225	0.200
	bipolar 3 layer	0.350	0.400	0.275	0.250	0.275	0.250

table 10: Best accuracies when different networks were trained with all 5 different intelligibility classes.